Genetic Epidemiology 21(Suppl 1): S760–S765 (2001)

Comparison of Likelihood Approaches for Combined Segregation and Linkage Analysis of a Complex Disease and a Candidate Gene Marker Under Different Ascertainment Schemes

Maria Martinez, Alisa M. Goldstein, and Jeffrey R. O'Connell

INSERM EMI 00-06 (M.M.), Paris, France; National Cancer Institute (A.M.G.), Bethesda, Maryland; University of Pittsburgh (J.R.O.), Pittsburgh, Pennsylvania

We compared two joint likelihood approaches, with complete (L1) or without (L2) linkage disequilibrium, under different ascertainment schemes, for the genetic analysis of the disease trait and marker gene 1 in replicate 42. Joint likelihoods were computed without a correction for the selection scheme. For the different sampling schemes we have explored, our results suggest that L1 is a more powerful approach than L2 to detect major gene and covariate effects as well as to identify accurately gene×covariate interaction effects in a common and complex disease such as the Genetic Analysis Workshop 12 MG6 simulated trait.

Key words: ascertainment, gene×covariate interactions, linkage disequilibrium, segregation and linkage analysis

INTRODUCTION

Combined segregation and linkage analysis is particularly appropriate for candidate marker genes that are tightly linked to the trait locus. Further, for an associated gene marker, genetic analyses can be conducted assuming complete linkage disequilibrium between trait and marker loci. The gene underlying the trait and the marker are thereby confounded. Such an approach is expected to increase both power and efficiency for testing and estimating genetic and covariate effects on trait variability. It has been applied to the genetic analysis of complex quantitative [Martinez et al., 1995] and qualitative [Goldstein et al., 2000] traits. When analyzing non-randomly selected families where the

Address reprint requests to Dr. Maria Martinez, INSERM EPI 00-06, Hôpital St. Louis, 1 Avenue Claude Vellefaux, 75010 Paris, France.

mode of ascertainment is not known, the validity of the parameter estimates and hypothesis testing using joint likelihoods are, however, questionable. To circumvent this problem, the conditional approach (mod score) has been advocated to be more appropriate [Risch, 1984]. Indeed, the mod score leads to unbiased genetic parameter estimates, provided the ascertainment scheme is based on trait only [Hodge et al., 1994]. This condition is not satisfied when the genetic marker is associated with the trait or is the trait gene itself. Our objective was to compare three (joint with or without complete linkage disequilibrium and conditional) likelihood approaches under different ascertainment schemes, for the genetic analysis of the disease trait and marker gene 1 in the Genetic Analysis Workshop (GAW) 12 simulated data set. However, in these data, we could not get an accurate maximum of the conditional likelihoods with gene 1. We thus reduced the comparisons to the two joint likelihood approaches. All analyses were conducted with the knowledge of the simulated model.

METHODS AND MODELS

Data set analyzed. We used the best replicate (i.e., 42) only. To reduce computation time, the 23 original pedigrees were split into the probands' family members and those including the family members of the probands' spouses. Deceased individuals had all unknown disease status and were considered to be unavailable for typing. A total of 46 reduced pedigrees (1,496 individuals, including 1,004 subjects alive and genotyped, out of whom 26% are affected) was thus obtained. In this sample, denoted Ped_Random, all but one pedigree had at least one affected member. We then used different sampling schemes to construct three other data sets. In each pedigree, one affected relative was assigned as the index case. In pedigrees with multiple affected subjects, the index case was randomly drawn from the nonfounder affected subjects. All first and second degree relatives of the index case were retained in the data set. Remaining sibships, including third and higher degree relatives, were retained if at least one relative was also affected. This data set, Ped_Index, contained a total of 45 pedigrees. The third data set was constructed by selecting in the Ped_Index sample, the pedigrees in which the index case had at least one secondary case within his/her first degree relatives (Ped_Mult sample, 23 pedigrees). Finally, we constructed a sample of nuclear families (parents and siblings) of the 45 index cases (Nuc_Fam). Table I shows the characteristics of the four constructed samples.

Statistical analyses to assess associated covariates. The effects of sex and gene 1 (sequence data on site 557) on disease and age of onset were assessed prior to our segregation and linkage analysis using the BMDP package. The non-ancestral and the ancestral sequence variants of gene 1 were coded as allele 1 and 2, respectively. In the

TABLE I. Characteristics of the Sample Data by Ascertainment Scheme

		No. subjects			No. subjects by family[a]	
	N	Total (aff)	Alive & genotyped (aff)	Rate of aff subjects	All	Alive & genotyped (aff)
Ped_Random	46	1,496 (260)	1,004 (260)	26%	32.5	21.8 (5.7)
Ped_Index	45	867 (232)	614 (232)	38%	19.3	13.6 (5.2)
Ped_Mult	23	533 (171)	392 (171)	44%	23.2	17.0 (7.4)
Nuc_Fam	45	238 (80)	175 (80)	46%	5.3	3.9 (1.8)

[a] Average family size(s).

Ped_Random sample, logistic regression analyses showed that both gender and presence of allele 2 were significantly associated with the disease ($p < 10^{-6}$): the probability of being affected increased in women and in carriers of allele 2. Product-limit estimation of disease survival function, using age of onset and age at exam for the affected and unaffected subjects, also showed significant differences in survival curves according to gender and gene 1 genotypes (Mantel-Cox, $p < 10^{-5}$). Linkage disequilibrium between disease and gene 1 was tested by means of the chi-square transmission/disequilibrium test (TDT) test [TDTLIKE program, Terwilliger, 1996]. In sibships with multiple affected sibs, we randomly chose one affected sib only. The available number of sibships (both parents genotyped and at least one heterozygous parent) was 33, 27, and 10 in Ped_Random and Ped_Index, Ped_Mult, and Nuc_Fam sample, respectively. Excess allele 2 transmission was significant in all (71%, $p = 2.8 \times 10^{-3}$ Ped_Random and Ped_Index; 69%, $p = 1.4 \times 10^{-2}$ Ped_Mult) but the Nuc_Fam (64%, $p = 0.14$) sample.

Combined segregation and linkage analysis of disease and gene 1. The joint transmission of the disease trait and the sequence data on site 557 (gene 1), while accounting for associated variables, was conducted using the regressive models [Bonney 1986, Bonney et al., 1988], and extended to take into account variable age of disease onset and censoring [Abel and Bonney, 1990], as implemented in the computer programs REGRESS [Demenais and Lathrop, 1994] and FINESSE [O'Connell et al., 1998]. The parameters of major gene effects are the frequency, q, of the disease allele (D), and the three genotype-specific baseline parameters (α_G). Family dependencies of unspecified origin (F.D.), of the i^{th} relative on his/her antecedents (j), are specified by Γji, the vector of regression coefficients [Demenais, 1991]. Hazard risks (period of follow up taken here from age 19 to 60), were modeled by means of a logistic function including α_G, Γj, a polynomial function of the logarithm of age, and sex as a covariate. The vector of regression coefficients of the age function ($\beta^k a$) and the regression coefficient for sex (βs) can be genotype-specific, i.e., $\beta^k a_G$ and/or βs_G. The linkage relationship between loci is specified with θ, the recombination fraction between the trait and marker locus. Gene 1 is located within the major gene directly affecting disease liability (MG6), thus combined segregation and linkage analysis was conducted by setting $\theta = 0$. When allowing for linkage disequilibrium (Δ) between trait and gene 1 loci, there are four haplotypes (d-1, d-2, D-1, D-2) with frequencies, h1, h2, h3, and h4, respectively. When assuming complete Δ (i.e., $h2 = h3 = 0$ and $h1 = 1-h4$), h4 was estimated along with all parameters of the combined segregation and linkage analysis. Joint likelihoods of the trait and gene 1 were computed in the four created data sets (without a correction for the ascertainment process) under two different approaches: assuming complete Δ (L1) and assuming linkage equilibrium (L2) between the trait and marker loci. When there is no Δ (L2), gene 1 allele frequencies in the analyzed data set were set to their maximum likelihood estimates, using ILINK from the VITESSE package [O'Connell and Weeks, 1995].

RESULTS AND DISCUSSION

Comparison of the two joint likelihood (L1 and L2) approaches in the random family sample (Ped_Random). A summary of the hypothesis testing in Ped_Random sample when maximizing joint likelihoods of the disease and gene 1 under complete linkage disequilibrium (L1) and under no linkage disequilibrium (L2) is presented in Table II. Family dependencies and covariate effects are detected under both L1 and L2,

TABLE II. Segregation and Linkage Analysis of Disease and Gene 1 in Ped_Random Sample When Complete Linkage Disequilibrium is Assumed (L1) or When Linkage Equilibrium is Assumed (L2)

A. χ^2 (p) values of genetic and covariates tests

Hypothesis	No F.D.	No MG	MG recessive	MG dominant	No sex effects	Linear age fct.
L1	51.5 ($< 10^{-9}$)	140.6 ($< 10^{-9}$)	80.9 ($< 10^{-9}$)	21.7 (310^{-6})	72.2 ($< 10^{-9}$)	67.1 ($< 10^{-9}$)
L2	53.0 ($< 10^{-9}$)	55.5 ($< 10^{-9}$)	13.5 (310^{-4})	16.6 (510^{-5})	73.4 ($< 10^{-9}$)	69.2 ($< 10^{-9}$)

B. p-values of MG×age, MG×sex interaction tests

Hypothesis	MG×sex		MG×age	
model	Age	MG × age	Sex	MG × sex
L1	0.054	0.053	0.018	0.018
L2	0.040	0.090	0.058	0.210

and with similar significance levels. In all analyses, females had greater disease risk than males, and hazard risk variations were best fitted by a polynomial rather than a linear function of age. L1 had, however, greater power to detect major gene effects than L2, and statistical significance for rejecting specific major gene effects (recessive or dominant) was also higher in L1 than L2. The main differences here were seen when testing for gene×covariate interaction effects. MG×age interactions were significant under L1 but not under L2, and whether or not MG×sex interactions were also included in the model: hazard variations with age were similar in 12 and 22 subjects and significantly different in 11 subjects. Conversely, significant MG×sex interactions were found under L2 only, and when no MG×age interactions were assumed: sex effects were higher in dd than in Dd or DD subjects. With L1, significance of sex×gene 1 interactions was borderline but, here, sex effects in 11 and 12 subjects were similar and higher than in 22 subjects.

From the answers, we know that affection status was defined through age and sex-specific thresholds on liability: disease risk increases with age, and is higher in females than in males. No gene×sex interaction effects were introduced into the simulated model. We also know that the true disease model is more complex than our analysis models. Indeed, MG6 (gene 1) accounts for a large part, but not all, of the liability variance. Also, liability does not explain all age at onset (AOO) variability. This variability is mainly dependent on MG7 genotypes. However, liability and AOO values are negatively correlated. Further, since MG7 and MG6 are tightly linked (1-cM distance between them), their genotypes (phenotypes) are not independently distributed within families. As a consequence, MG6×age interaction effects are likely to exist. Our analyses in the non-selected sample (Ped_Random) showed that both likelihood approaches provided similar conclusions: sex and major gene effects were highly significant. These effects, along with additional family dependencies, were shown to fit familial transmission of the disease. These conclusions are in agreement with the simulated MG6 trait. The two likelihood approaches did not provide, however, the same estimated values of the major gene (i.e., susceptibility allele frequency, penetrances). The answers provided MG6 disease allele frequency (0.232), but not the age and sex-specific penetrance values. The disease allele frequency was almost unbiased with L1 (h4 = 0.233), and slightly overestimated with L2 (q = 0.262). The two approaches also differ with respect to gene×covariate interaction tests. Gene 1×age effects were significant with L1 but not with L2. Conversely, MG×sex interaction effects were significant with L2 only. Overall, in these data, the phenotype and genetic variations were more accurately estimated with L1 than L2.

Combined segregation and linkage analysis under different ascertainment schemes. Our sampling schemes increased the proportion of affected cases in the selected samples but they also varied the number of families (N), and led to a reduction in the sizes (S) of the ascertained families (Table I). To evaluate the effect of a sampling scheme under L1 and L2, we divided the total χ^2 values by the sample size of subjects (i.e., N*S). This allowed us to derive a measure of the contribution (i.e., amount of information) of a family of the same size within each constructed sample. Table III reports the average contribution of a family of size five and the summary of the segregation-linkage results. Significance for family dependencies ($p \leq 3 \times 10^{-9}$), sex effects ($p \leq 3 \times 10^{-3}$) and non-linear age risks ($p \leq 3 \times 10^{-5}$) was again obtained in all "selected" samples. For these three tests, the average chi-square contributions are slightly higher with L2 than L1, and whichever the selection scheme. Within the pedigree samples, the amount of information is also slightly increased in the "selected" samples.

Major gene effects (i.e., rejection of no MG) were significant in all samples ($p \leq 9 \times 10^{-5}$) except for L2 in the Nuc_Fam sample ($p = 0.16$). Table III shows that, whichever the sampling scheme, the average contribution of a family for examining MG effects is much greater with L1 than L2, especially in the Nuc_Fam sample, where it is 5.6 times higher with L1 than L2. The contribution is highest in the Ped_Mult sample, and lowest in the Nuc_Fam sample, especially with L2. Thus, a pedigree of 25 members is much more informative than five nuclear families of five members each, with L2 but not with L1. These results suggest that the power to detect major gene effects is only slightly affected by the ascertainment scheme with L1 but not with L2. Linkage disequilibrium between disease and gene 1 allele 2 was highly significant in all samples ($p \leq 10^{-9}$ in Ped_Random, Ped_Index and Ped_Mult; $p = 1.1 \times 10^{-4}$ in Nuc_fam). Note, however, that the relative contribution of a family (results not shown) is higher in Nuc_Fam (0.37) than in Ped_Random (0.31). These results confirm the better power of joint likelihood approaches to detect linkage disequilibrium over nonparametric tests, such as the TDT.

TABLE III. Segregation and Linkage Analysis of Disease and Gene 1: Average χ^2 Values for a Family of Size 5 and Parameter Estimates by Sampling Scheme and When Complete Linkage Disequilibrium is Assumed (L1) or When Linkage Equilibrium is Assumed (L2)

							Best model:	
					MG×covariate		parameter estimates	
	No major	No	No sex	Linear	MG	MG		Lifetime
	gene	F.D.	effects	age fct.	×sex	×age	f(D)	prev. (M/F)
L1								
Ped_Random	0.47[a]	0.17[a]	0.24[a]	0.22[a]	*0.01*	0.02[d]	0.233	0.22/0.48
Ped_Index	0.47[a]	0.32[a]	0.27[a]	0.37[a]	*0.02*	0.05[c]	0.300	0.32/0.59
Ped_Mult	0.48[a]	0.41[a]	0.30[a]	0.34[a]	*0.02*	0.08[c]	0.320	0.37/0.63
Nuc_Fam	0.45[b]	0.72[a]	0.19[c]	0.38[b]	0.13[d]	*0.05*	0.375	0.33/0.62
L2								
Ped_Random	0.19[a]	0.18[a]	0.25[a]	0.23[a]	0.02[d]	*0.01*	0.262	0.22/0.52
Ped_Index	0.18[a]	0.36[a]	0.30[a]	0.39[a]	*0.02*	0.05[d]	0.489	0.34/0.72
Ped_Mult	0.20[b]	0.47[a]	0.35[a]	0.38[a]	*0.01*	*0.01*	0.336	0.39/0.73
Nuc_Fam	*0.08*	0.74[a]	0.19[c]	0.27[c]	*0.01*	*0.06*	--	--

[a] $p < 10^{-7}$

[b] $10^{-7} < p < 10^{-4}$

[c] $10^{-4} < p < 10^{-2}$

[d] $10^{-2} < p < 0.05$

[e] Significant χ^2 tests (shown in italics).

With L2, MG×sex interaction effects were found to be significant in the Ped_Random sample only, whereas significant MG×age effects were detected in the Ped_Index sample only. With L1, these tests led to consistent conclusions (i.e., significant gene 1×age and nonsignificant gene 1×sex interactions) in the different samples except in the Nuc_Fam sample, where gene 1×age effects were not significant ($p = 0.17$), but gene 1×sex effects were significant ($p = 0.013$). Thus, to prevent making false inferences about an "unlikely" gene 1×sex interaction with L1, one pedigree of 25 members is more informative than five nuclear families of five members. Not surprisingly, the disease allele and lifetime prevalence estimates were overestimated in the selected samples. In the pedigree samples, biases increased as the selection criteria increased, but were lower than in the Nuc_Fam sample. Within a specific selection scheme, these biases were higher under L2 than L1.

Altogether, for the different selection schemes we have explored in these data, L1 appeared to be a more powerful joint likelihood approach than L2 in detecting genetic and covariate effects as well as accurately identifing gene×covariate interaction effects in a disease such as the GAW12 MG6 simulated trait. Here, we used the average family χ^2 contributions to assess the relative power of L1 and L2. Further simulation studies will be conducted to measure formally the power of these two approaches as well as to extend this investigation to other candidate gene and gene×covariate interaction models.

ACKNOWLEDGMENTS

This work was supported by a BIOMED grant (BMH4-CT97-2532)

REFERENCES

Abel L, Bonney GE. 1990. A time-dependent logistic hazard function for modeling variable age of onset in analysis of familial diseases. Genet Epidemiol 7:391-07.

Bonney GE. 1986. Regression logistic models for familial disease and other binary traits. Biometrics 42:611-25.

Bonney GE, Lathrop GM, Lalouel JM. 1988. Combined linkage and segregation analysis using regressive models. Am J Hum Genet 43:29-37.

Demenais F. 1991. Regressive logistic models for familial diseases: A formulation assuming an underlying liability model. Am J Hum Genet 49:773-85.

Demenais F, Lathrop GM. 1994. REGRESS: A computer program including the regressive approach into the LINKAGE program. Genet Epidemiol 11:A291.

Goldstein AM, Martinez M, Tucker MA, et al. 2000. Gene-covariate interaction between dysplatic nevi and the CDKN2A gene in American melanoma-prone families. Cancer Epidemiol Biomarkers Prev 9:889-94.

Hodge SE, Elston RC. 1994. Lods, Wrods, and Mods: The interpretation of lod scores calculated under different models. Genet Epidemiol 11:329-42.

Martinez M, Abel L, Demenais F. 1995. How can maximum likelihood methods reveal candidate gene effects on a quantitative trait? Genet Epidemiol 12:789-94.

O'Connell JR on behalf of the European group. 1998. Dissecting complex diseases with FINESSE. Genet Epidemiol 15:A521.

O'Connell JR, Weeks DE. 1995. The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nat Genet 11:402-8.

Risch N. 1984. Segregation analysis incorporating linkage markers. I. Single-locus models with an application to type I diabetes. Am J Hum Genet 36:363-86.

Terwilliger J. 1996. Program ANALYZE. ftp://linkage.cpmc.columbia.edu.